BUILDING CHANGE DETECTION BY W-SHAPE RESUNET++ NETWORK WITH TRIPLE ATTENTION MECHANISM

A. Eftekhari [1*], F. Samadzadegan [1], F. Dadrass Javan [1, 2]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran 1439957131, Iran - (a.eftekhari, samadz, fdadrasjavan)@ut.ac.ir
[2] Faculty of Geo-Information Science and Earth Observation (ITC), University of Twente, 7522 NB Enschede, the Netherlands - f.dadrassjavan@utwente.nl

Commission IV, WG IV/3

KEY WORDS: Remote Sensing Image Change Detection, Deep Learning, Attention Mechanism, W-shape Networks, High-resolution Images, Dual Loss Function

ABSTRACT:

Building change detection in high resolution remote sensing images is one of the most important and applied topics in urban management and urban planning. Different environmental illumination conditions and registration problem are the most error resource in the bitemporal images that will cause pseudochanges in results. On the other hand, the use of deep learning technologies especially convolutional neural networks (CNNs) has been successful and considered, but usually causes the loss of shape and detail at the edges. Accordingly, we propose a W-shape ResUnet++ network in which images with different environmental conditions enter the network independently. ResUnet++ is a network with residual blocks, triple attention blocks and Atrous Spatial Pyramidal Pooling. ResUnet++ is used on both sides of the network to extract deeper and discriminator features. This improves the channel and spatial inter-dependencies, while at the same time reducing the computational cost. After that, the Euclidean distance between the features is computed and the deconvolution is done. Also, a dual loss function is designed that used the weighted binary cross entropy to solve the unbalance between the changed and unchanged data in change detection training data and in the second part, we used the mask–boundary consistency constraints that the condition of converging the edges of the training data and the predicted edge in the loss function has been added. We implemented the proposed method on two remote sensing datasets and then compared the results with state-of-the-art methods. The *F1* score improved 1.52 % and 4.22 % by using the proposed model in the first and second dataset, respectively.

1. TNTRODUCTION

Building change detection is one of the important applications of remote sensing images, which means detecting and extracting spatial changes of buildings in a geographical entity from multi-temporal imagery (Khelifi and Mignotte, 2020). Building change detection research has an important role in urban management, urban planning, crisis management, damage assessment and updating topographic and cadastre maps (Zheng et al., 2021). With the advancement of satellite data collection technology, many high-resolution images have become available. In these images, building objects have more space and shape features; therefore, high-resolution images became a good source for buildings change detection (Chen et al., 2021).

In recent years, the use of automatic methods for change detection has grown a lot, and in the meantime, methods based on machine learning and especially deep learning (DL) have been the focus of researchers. Convolution neural network (CNN) is one of the DL methods for image data processing which is used to extract extracting high-level discriminatory features. A fully convolutional network (FCN) has been successfully applied to the end-to-end change detection. U-Net is a one of the standard FCNs used in change detection research (Song et al., 2021). Recently, an improved UNet++ framework was introduced for very high resolution (VHR) images change detection (Peng et al., 2019). The ResUNet (Jha et al., 2019) network is a form of U-Net architecture provided competitive results for segmentation. Therefore, we consider this network as the basis for our proposed method.

DL methods used to changes detection are divided into two groups: classification based and metric based (Khelifi and Mignotte, 2020). In classification based methods, the first and second time images are entered into a single-branch deep network as a joint vector and deep features are extracted from the bi-temporal images. In metric methods, a dual-branch network is used to extract the amount of change by comparing the distance parameter between two bi-temporal images. W-shape network is a kind of dual-branch network in which images are entered from both sides of the network and deep features are extracted from each image separately (Zhang et al., 2021). Then by calculating the Euclidean distance between the features, network upsampling is done as a single branch. Unlike Siamese networks, weights are not shared in these networks.

In deep networks, extracting discriminative features from images is very important to achieve better results and is still one of the challenging issues. The use of attention mechanisms is one of the ways of extracting distinguishing features in deep network. Channel attention (Liu et al., 2020), spatial attention or both of them (Ding et al., 2021; Ma et al., 2022) are used to improve the performance of CNNs. These attention mechanisms have improved representation of the features generated by standard convolutional layers. Using learning attention weights increases the ability to learn where to attend and focus more on the target objects.

Squeeze and- excitation networks (SENet) are one of the most used methods (Guo et al., 2022) which modeled channel interrelationships in feature maps by learning the modulation weights of each channel. SENet gives good performance with very low computational cost. Convolutional Block Attention Module (CBAM) and Bottleneck Attention Module (BAM) have used SENet successfully. Self-attention mechanism is another kind of attention module encoded a long range of local

features as background information, thus increasing the quality of the represented features. Spatial Attention Mechanism (SAM) and Channel Attention Mechanism (CAM) are two types of self-attention methods have been successful in simulating long-range dependence(Chen et al., 2021). Considering that the previously attention mentioned methods require an extra learnable, our goal in this article is to use triple attention mechanism which is cheap but effective attention method with equal or even better performance than the previous ones. Triple attention gets cross-interaction by calculating attention weights to give rich feature representation.

Since the shape details are lost in the expansion process of U-shaped networks, we propose a dual loss function that combines mask boundary consistency constraints (MBCC) and weighted binary cross entropy (WBCE) loss function to solve mentioned problem. The MBCC loss design to minimize the difference between the boundaries derived from the building changed map and the boundaries derived from predicted change map. By Using MBCC-BCE loss function, the network performance in detecting change, especially at the boundary of changed buildings is increased.

In this article, we propose the W-shape ResUNet++ network by triple attention for building change detection. The proposed method was implemented in the building change detection datasets (BCDD) (Ji et al., 2019) and LEVIR-CD (Chen and Shi, 2020), In addition, several state-of-the-art methods have been used for evaluating proposed network performance. In summary, the article includes the following:

1) We suggest the W-shape ResUNet++ architecture, which is change detection network that takes advantage of residual blocks, triple attention blocks and Atrous Spatial Pyramidal Pooling (ASPP). W-shape ResUNet++ by MBCC-BCE loss function improved the building change detection results significantly compared to other state-of-the-art methods.

2) We implemented our proposed method in BCDD and LEVIR-CD. In addition, it was compared with several state-of-the-art methods.

2. METHODOLOGY

2.1 Network Overview

The architecture of proposed method is shown in Figure 1. The proposed method is based on a W-shape ResUnet++ network with a triple attention mechanism. ResUnet++ is a network with residual blocks, triple attention blocks and ASPP. ResUnet++ is used on both sides of the network to extract deeper and discriminator features. This improves the channel and spatial dependencies, while reducing the computational time at the same time. After that, the Euclidean distance between the features is computed and the de-convolution is done. The skip connection strategy from the W-shape network is generalized on both sides. So that the information of low-level features is copied symmetrically from both sides of the network in the expansion path and is combined and stacked with high-level information. Then convolution and batch normalization is performed, and ASPP is applied before sigmoid activation layer. It is noteworthy that, the contextual information is extracted at different scales by using ASPP and, it captures appropriate multi-scale information for change detection tasks.

2.2 Triple Attention

Triple attention is a new attention weights computing method that is captured cross-dimension interplay by using a three-branch design whose architecture is given in figure 2. Triple attention creates inter-dimensional connections by using rotation action came after by residual transformations for an input tensor, and encodes cross-channel and spatial information at very low computational cost. (Huang et al., 2021). In triple attention, the input feature (x) with (C×H×W) dimension is in charge of summing cross-interaction features between both the spatial (H or W) dimension and the channel dimension C. The input vectors is permuting in each direction and, then transmitting the vector via a *Z-pool,* followed by a k × k convolution layer.

The Z-pool layer concatenates the average and max pooled features among that dimension to reduce the zeros dimension. This enables the layer to provide a better representation of the true tensor while simultaneously reducing its depth to significantly reduce computation. Its mathematical formula is given below.

$$\mathrm{Z-pool(x)} = [\mathrm{MaxPool_{0d}(x), AvgPool_{0d}(x)}] \quad (1)$$

where $0d$ = 0th-dimension across
$\mathrm{MaxPool_{0d}(x)}$= max pooling operator
$\mathrm{AvgPool_{0d}(x)}$= average pooling operator

$0d$ is the 0th-dimension across which the max and average pooling operations happen. For example, the Z-Pool of a tensor with (C × H × W) dimension is equal to a tensor with (2 × H ×W) dimension.

In the last branch, a standard convolution layer by k × k kernel size and a batch normalization layer are passed over the tensor with the shape (2 × H × W). Then, a sigmoid activation layer (σ) is used to calculate the shape attention weight (1 × H × W) that is applied to the input x. A simple averaging function is used to sum the refined tensors by (C × H × W) shape generated by each of the three branches.

2.3 Atrous Spatial Pyramidal Pooling

ASPP gets from spatial pyramidal pooling and using for re-sampling features at multiple scales. ASPP model the contextual representation and effectively expands received contexts while keeping spatial resolution (Chen et al., 2018). Many parallel atrous convolutions layers with different rates are fused to take the contextual information at various scales. Atrous convolution expands the field of view to capture multi-scale features and also generates more parameters (Li et al., 2021). In the proposed algorithm, ASPP operates as a bridge between encoder and decoder in the both side of networks, as shown in Figure 1. ASPP model has shown promising results by providing multi-scale information. Hence, we use ASPP to extract useful multi-scale information for the task of building change detection.

2.4 MBCC-WBCE dual loss function

MBCC: Boundaries encode important shape information and can be used to improve shape information. Here, we used the boundary of the change mask. Thus, we defined an appropriate loss function by extracting the boundary of the changed and unchanged regions and defining the consistency constraints between the boundary and the change mask. Therefore, to minimizing the difference between the ground truth change mask and the predicted change mask, the difference between the boundaries extracted from the change and the predicted change

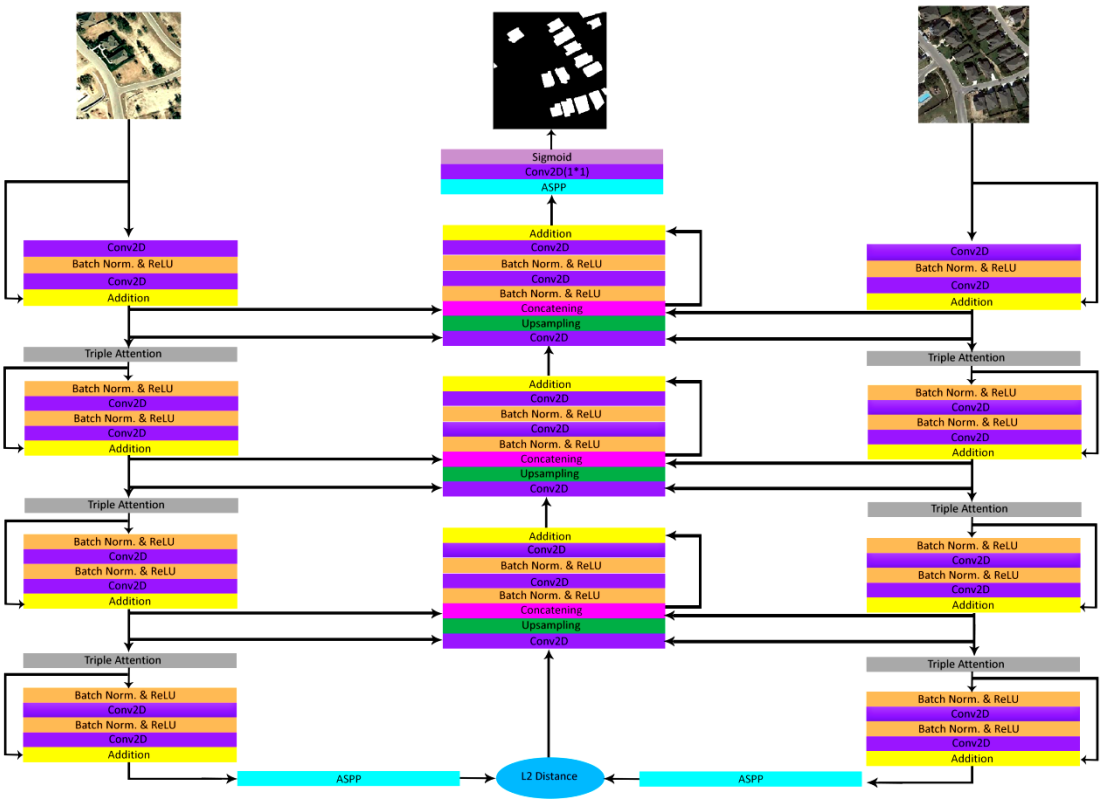

Figure 1. Block diagram of the proposed W-shape ResUNet++ architecture

mask must also be minimized in the process of network training. In order to model the consistency constraints between the change boundary and the change mask, the change mask is converted to the change boundary using different functions. The transformed function calculates the maximum difference between the predicted change mask and its neighboring pixels, so that the values of the change boundaries are assigned to 1 and the other values are assigned to 0 for the transformed boundary map. The transformed function is obtained from Equation (2):

$$f_{m\to b}(x) = -\text{maxPooling}(x; \text{kernel size} = 3. \text{stride} = 1 \quad (2)$$

where x= the probability values of changes in the predicted change area

This function is also applied in training process. The results of a number of boundaries extracted from the change mask are shown in the figure below, in which the function $f_{m\to b}(x)$is used to convert the change mask to the change boundary. To apply consistency constrains between the mask and the change boundary, the loss function of the change MBCC is defined to minimize the difference between them:

$$L_{mbc}(M_{pre}) = \left|f_{m\to b}(M_{pre}) - B_{pre}\right| = \left|f_{m\to b}(M_{pre}) - f_{m\to b}(M_{gt})\right| \quad (3)$$

where |.| = distance function as L1
Mpre= the mask of predicted changes
Bpre = the boundary of predicted changes
Mgt = the mask of ground truth changes

WBCE: One of the problems with building change detection training data is the imbalance between changed and unchanged data. Usually the unchanged data is multiplied many times the changed data. In the case of simple loss functions, this imbalance is not taken into account. Therefore, we have used the WBCE loss function for the problem of imbalance and its effect on final accuracy as Equation (4). So, its weights are determined in proportion to the number of training data.

$$L_{wbce}(M_{pre}) = -\sum_{i=1}^{m}(w_1 Mgt_i \times \log(Mpre_i) + w_2(1 - Mgt_i) \times \log(1 - Mpre_i)) \quad (4)$$

where w_1= weight for changed pixels
w_2= weight for unchanged pixels
Mpre= the mask of predicted changes
Mgt = the mask of ground truth changes

Final loss: Final value of the loss function is obtained from the sum of the MBCC which is introduced in Equation (3) and WBCE which is introduced in Equation (4). Thus, the final loss function is obtained from the following Equation:

$$L_f = \lambda_1 L_{mbc}(M_{pre}) + \lambda_2 L_{wbce}(M_{pre}) \quad (5)$$

where λ1, λ2 = coefficients for combining two losses

λ1 and λ2 are determined in experiments.

2.5 Accuracy Assessment

Precision-Recall is an appropriate measure to evaluate results when classes are unbalanced. Since in the problem of change detection, usually the number of changed points is much less than unchanged and there is no balance between these two data, the use Precision-Recall metrics is appropriate. So, we use the precision (Pr), recall (Re) and F1-score (F1) for evaluating results(Tharwat, 2018).

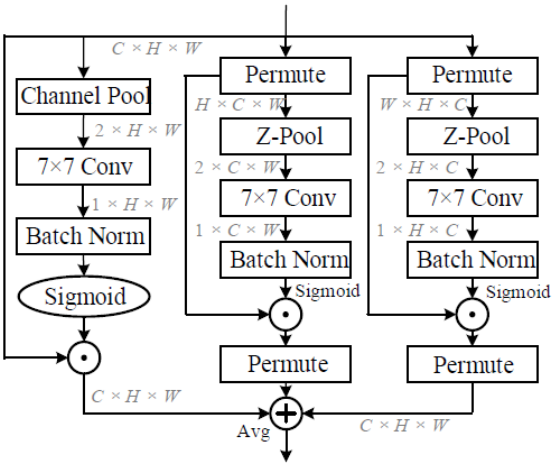

Figure 2. Triple attention architecture

3. RESULTAS AND DISCUSSION

In this section, we show the results of implementing the proposed method on two datasets. We also select some state-of-the-art to evaluate the proposed method and compare and analyze the proposed network results with them. The Tensorflow backend is used to implement the proposed method, in which a single Tesla P100-PCIE-16GB. The Adam optimizer was developed with learning rates decaying from 1e-1 to 1e-4, and the training data with a batch size of 20 entered the network. The λ1 value was set to 1, and the value of λ2 was set to 0.5 based on the performance experiments. 70% of the data was used for training, 20% for testing and 10% for validation in both the BCDD and LEVIR-CD datasets. Due to the graphics card's limitations, we chose an input size of 128 ×128 pixels.

3.1 Results on LEVIR-CD Dataset

Hao Chen and Zhenwei Shi introduced LEVIR-CD as an open-source building change detection dataset. It consists of 637 very high resolution images (0.5 meters/pixel) with patch sizes of 1024 in1024 pixels from Google Earth images (Chen and Shi, 2020). The images were taken between 2002 to 2018 and are from different cities in Texas in the US. The labeled LEVIR-CD contains 31,333 separate building changes, with an average of 50 buildings changed in 1024 by 1024 images. Most of the changes are new buildings constructed with approximately 987 pixels per image.

The results of implementing the proposed method on LEVIR-CD are given in Table 1. Here, Unet (Resnet50) (Diakogiannis et al., 2020) is used as a base model, and we have investigated the effect of using the ResUnet++, W-shape models and the proposed dual loss function. The first point in Table.1 is the effect of using ResUnet++ network, which increases F1 by 1.4% in the case of binary cross entropy loss and 3.44% in the case of using dual loss function. Also Using W-shape structure increases the F1 by 6.94 % on average. The use of proposed loss function has increased F1-score value by an average of 3.79 % in all three cases of Table. 1. In particular, the new loss function has further increased the value of Recall, which indicates a decrease in FNs due to the use of this function.

3.2 Results on BCDD Dataset

BCDD is a dataset for building change detection from an area in New Zealand where an earthquake with 6.3 magnitudes happen in February 2011. Then, aerial images of 2012 after the earthquake and images after reconstruction in 2016 were taken. The two 100% overlapped datasets by 20 cm spatial resolution consist of 12796 buildings in 20.5 km2 and 16077 buildings in the same area in the 2016 data (Ji et al., 2019).

The results of implementing the proposed method on BCDD are given in Table 2. Here, Unet (Resnet50) is used as a base model, and we have investigated the effect of using the ResUnet++, W-shape models and the proposed dual loss function. The results in Table 2 also show that the use of ResUnet++ architecture compared to Unet has increased the accuracy and the value of the F1 parameter has increased by an average of 3.13%. Also Using W-shape structure increases the F1 by 4.72 % on average. The use of proposed loss function has increased F1-score value by an average of 2.47 % in all three cases of Table. 2. Also, the value of precision has increased significantly, indicating a suitable decrease in FP due to the use of the dual loss function.

	Binary cross entropy loss			MBCC-WBC loss		
Methods	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
Unet-Resnet50	81.18	83.28	82.22	82.56	84.93	83.73
ResUnet ++	82.57	84.19	83.37	87.95	85.31	86.61
W-shape ResUnet ++	89.73	87.12	88.41	94.28	92.53	93.40

Table 1. Ablation study of W-shape ResUnet++ and dual loss based on MBCC-WBCE loss on LEVIR-CD validation set.

	Binary cross entropy loss			MBCC-WBC loss		
Methods	Precision (%)	Recall (%)	F1-Score (%)	Precision (%)	Recall (%)	F1-Score (%)
Unet-Resnet50	84.21	84.57	84.39	86.19	85.81	86.00
ResUnet ++	86.51	87.53	87.02	89.17	88.26	88.71
W-shape ResUnet ++	91.48	89.37	90.41	95.62	91.71	93.62

Table 2. Ablation study of W-shape ResUnet++ and dual loss based on MBCC-WBCE loss on BCDD validation set.

3.3 Comparisons on state-of-the-art methods

To better understand the results of the proposed method and see how well it worked, we implemented the other state-of-the-art CD methods on both datasets and compared them with the proposed method statistically and visually in this part. Three state-of-the-art methods are implemented to assess the performance of the proposed method. FC-EF (Boulch, 2018) is the method using fully convolutional Siamese networks for change detection. STANet (Chen and Shi, 2020) is a method which was proposed by creator of LEVIR-CD dataset. They modeled the spatial-temporal relationships by using self-attention mechanism. AGCDetNet (Song and Jiang, 2021) is attention-guided end-to-end change detection network in which multilevel features and multi-scale context are enhance by using spatial attention and channel-wise attention-guided interference filtering unit module.

The visual results of comparing the proposed method with the state-of-the-art methods have been brought in Figure 3 for LEVIR-CD dataset and Figure 4 for BCDD dataset. The Pr, Re and F1 values of the three methods and the proposed method are also compared for BCDD dataset graphically in Figure 5.

According to the results, the proposed method in building change detection has a superior performance than the previous appropriate change detection methods. Therefore, the F1 value

has grown by 4.22% on the BCDD dataset compared to the best results of the state-of-the-art methods.

The visual results shown in Figures 3 and 4 show that the proposed method has detected building changes very well. Using triple attention and ASPP in the proposed method improves performance of it and changes are identified in good detail. While in the state-of-the-art methods, changes are not fully identified or additional changes are placed in some parts. Also, in the results of the proposed method, the edges are extracted much better than the rest of the methods, which indicates the suitable operator of the proposed loss function.

balanced the effect of the inequality between changed area and unchanged area on the network, as well as better extraction of edges in buildings. The effect of using ResUnet++, w-shape network, and the proposed loss function were shown step by step, and It was proved that the use of our method improves the accuracy of the evaluation metrics in the two LEVIR-CD and BCDD datasets. Finally, the proposed method was compared with three mentioned methods and the results were presented quantitatively and visually.

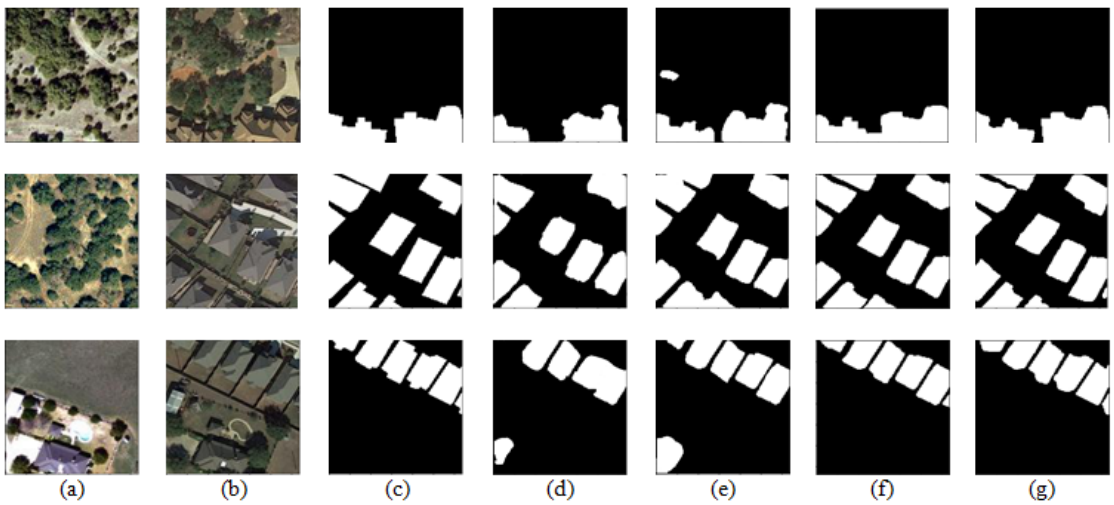

Figure 3. Visual comparison between state-of-the-art methods and proposed method on the LEVIR-CD dataset. (a) unchanged image, (b) changed image, (c) label, (d) FC-EF, (e) STANet, (g) AGCDetNet , i: proposed method.

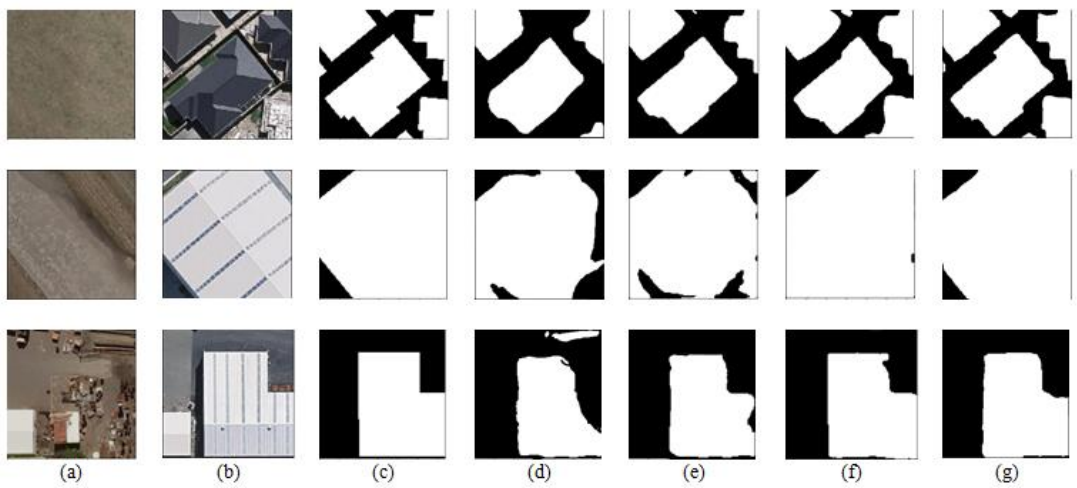

Figure 4. Visual comparison between state-of-the-art methods and proposed method on the BCDD dataset. (a) unchanged image, (b) changed image, (c) label, (d) FC-EF, (e) STANet, (g) AGCDetNet , i: proposed method.

4. CONCLUSIONS

In this article, we proposed a W-shape ResUnet++ network with triple attention and ASPP algorithms for high resolution building change detection. The use of a triple attention mechanism allows to the extraction of suitable discriminator properties to detect changes in more detail at an optimal time. We also used the MBCC-WBCE dual loss function, which

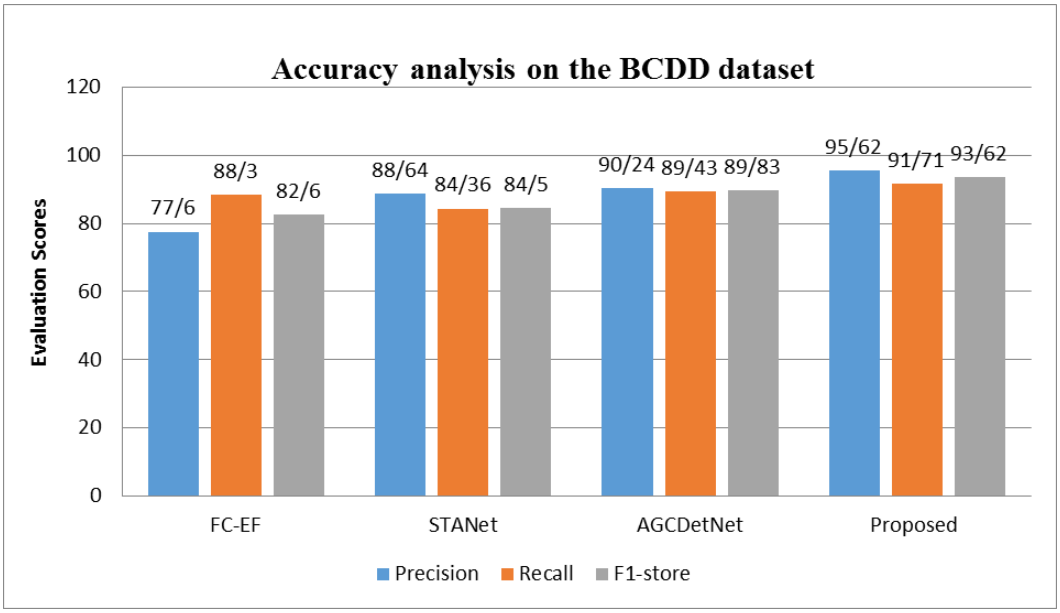

Figure 5. Accuracy comparison between state-of-the-art and proposed methods.

REFERENCES

Boulch, R.C.D.L.S., 2018. Fully convolutional siamese networks for change detection, in: Proceedings - International Conference on Image Processing, ICIP (2018).

Chen, H., Shi, Z., 2020. A spatial-temporal attention-based method and a new dataset for remote sensing image change detection. Remote Sens.

Chen, J., Yuan, Z., Peng, J., Chen, L., Huang, H., Zhu, J., Liu, Y., Li, H., 2021. DASNet: Dual Attentive Fully Convolutional Siamese Networks for Change Detection in High-Resolution Satellite Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14, 219-228

Chen, L.C., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L., 2018. DeepLab: Semantic Image Segmentation with Deep Convolutional Nets, Atrous Convolution, and Fully Connected CRFs. IEEE Trans. Pattern Anal. Mach. Intell.

Diakogiannis, F.I., Waldner, F., Caccetta, P., Wu, C., 2020. ResUNet-a: A deep learning framework for semantic segmentation of remotely sensed data. ISPRS J. Photogramm. Remote Sens. 162.

Ding, Q., Shao, Z., Huang, X., Altan, O., 2021. DSA-Net: A novel deeply supervised attention-guided network for building change detection in high-resolution remote sensing images. Int. J. Appl. Earth Obs. Geoinf. 105.

Guo, M.H., Xu, T.X., Liu, J.J., Liu, Z.N., Jiang, P.T., Mu, T.J., Zhang, S.H., Martin, R.R., Cheng, M.M., Hu, S.M.,2022. Attention mechanisms in computer vision: A survey. Comput. Vis. Media.

Jha, D., Smedsrud, P.H., Riegler, M.A., Johansen, D., De Lange, T., Halvorsen, P., Johansen, H.D., 2019. ResUNet++: An Advanced Architecture for Medical Image Segmentation, in: Proceedings - 2019 IEEE International Symposium on Multimedia, ISM 2019.

Ji, S., Wei, S., Lu, M., 2019. Fully Convolutional Networks for Multisource Building Extraction from an Open Aerial and Satellite Imagery Data Set. IEEE Trans. Geosci. Remote Sens. 57.

Khelifi, L., Mignotte, M., 2020. Deep Learning for Change Detection in Remote Sensing Images: Comprehensive Review and Meta-Analysis. IEEE Access.

Li, Y.C., Li, H.C., Hu, W.S., Yu, H.L., 2021. DSPCANet: Dual-Channel Scale-Aware Segmentation Network with Position and Channel Attentions for High-Resolution Aerial Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14.

Liu, J.J., Hou, Q., Cheng, M.M., Wang, C., Feng, J., 2020. Improving convolutional networks with self-calibrated convolutions, in: Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition.

Ma, J., Shi, G., Li, Y., Zhao, Z., 2022. MAFF-Net: Multi-Attention Guided Feature Fusion Network for Change Detection in Remote Sensing Images. Sensors 22.

Peng, D., Zhang, Y., Guan, H., 2019. End-to-end change detection for high resolution satellite images using improved UNet++. Remote Sens. 11.

Song, K., Jiang, J., 2021. AGCDetNet:An Attention-Guided Network for Building Change Detection in High-Resolution Remote Sensing Images. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. 14.

Song, L., Xia, M., Jin, J., Qian, M., Zhang, Y., 2021. SUACDNet: Attentional change detection network based on siamese U-shaped structure. Int. J. Appl. Earth Obs. Geoinf. 105.

Tharwat, A., 2018. Classification assessment methods. Appl. Comput. Informatics 17.

Zhang, H., Wang, M., Wang, F., Yang, G., Zhang, Y., Jia, J., Wang, S., 2021. A novel squeeze-and-excitation W-net for 2D and 3D building change detection with multi-source and multi-feature remote sensing data. Remote Sens.

Zheng, Z., Wan, Y., Zhang, Y., Xiang, S., Peng, D., Zhang, B., 2021. CLNet: Cross-layer convolutional neural network for change detection in optical remote sensing imagery. ISPRS J. Photogramm. Remote Sens. 175.